残差と近似解の誤差 (数値計算のアルゴリズムの研究)

AUTHOR(S):

平野, 菅保

CITATION:

平野, 菅保. 残差と近似解の誤差 (数値計算のアルゴリズムの研究). 数理解析研究所講究録 1981, 422: 59-77

ISSUE DATE:

1981-04

URL:

http://hdl.handle.net/2433/102562

数理解析研究所講究録
422巻 1981年 59-77

残差と近似解の誤差

東芝　　平野 菅保

I　有限桁の係数と近似解

実際に電子計算機を使用して数値計算をする場合、n元連立1次方程式

$$\sum_{j=1}^{n} a_{Tij} \cdot x_j - C_{Ti} = 0 \qquad i = 1, 2, \cdots, n \qquad (1)$$

a_{Tij}, C_{Ti} ：実数

の係数a_{Tij}，定数C_{Ti}は、0を除くと、一般には実数で与えられず、0でない係数a_{Tij}，定数C_{Ti}は、それぞれ有限桁の数値で与えられるので、それぞれ与えられる最下位の桁の次の桁から下位の桁の数値は、すべて不明である。したがって、不明の桁をすべて0として得られる係数a_{ij}，定数C_iを持つn元連立1次方程式

$$\sum_{j=1}^{n} a_{ij} \cdot x_j - C_i = 0 \qquad i = 1, 2, \cdots, n \qquad (2)$$

a_{ij}, C_i：実数

がn元連立1次方程式(1)を代表しているとして、n元連立

1次方程式(2)を解くことになる。ゆえに、n元連立1次方程式(2)の0でない係数a_{ij}, 定数C_iは

$$\begin{aligned} a_{Tij} &= a_{ij} + \varepsilon_{Tij} \qquad -\Delta a_{ij} \leqq \varepsilon_{Tij} \leqq \Delta a_{ij} \\ C_{Ti} &= C_i + \varepsilon_{Ti} \qquad -\Delta C_i \leqq \varepsilon_{Ti} \leqq \Delta C_i \\ \Delta a_{ij} &> 0 \qquad\qquad \Delta C_i > 0 \end{aligned} \tag{3}$$

Δa_{ij}：0でない係数a_{ij}の与えられる最下位の桁の単位の1/2

ΔC_i：0でない定数C_iの与えられる最下位の桁の単位の1/2

を満足することのみ既知であるn元連立1次方程式(1)の係数a_{Tij}, 定数C_{Ti}を代表していなければならない。すなわち

$$-\Delta a_{ij} \leqq \varepsilon_{ij} \leqq \Delta a_{ij} \qquad -\Delta C_i \leqq \varepsilon_i \leqq \Delta C_i \tag{4}$$

を満足する任意の誤差ε_{ij}, ε_iを実数a_{ij}, C_iに加えた実数$(a_{ij}+\varepsilon_{ij})$, $(C_i+\varepsilon_i)$を係数、定数に持つn元連立1次方程式

$$\sum_{j=1}^{n}(a_{ij}+\varepsilon_{ij})\cdot x_j - (C_i+\varepsilon_i) = 0 \tag{5}$$

$$i = 1, 2, \cdots, n$$

の解x_j $(j=1,2\cdots,n)$よりも、n元連立1次方程式(1)の解に近接していると認めることのできる近似解を、n元連立1次方程式(2)から求めることは一般にはできない。したがって、その近似解x_j $(j=1,2,\cdots,n)$をn元連立1次方程式(2)

を満足する近似解とする。

Ⅱ　残差と近似解

n 元連立1次方程式(5)の解 $\bar{x}_j$ $(j=1,2,\cdots,n)$ を n 元連立1次方程式(2)に代入すると、

$$\sum_{j=1}^{n} a_{ij}\cdot\bar{x}_j - c_i = -\sum_{j=1}^{n}\varepsilon_{ij}\cdot\bar{x}_j + \varepsilon_i \qquad (6)$$

$$i=1,2,\cdots,n$$

となり、すべての残差（右辺）が0になるとは限らない。すなわち、n 元連立1次方程式(2)の係数 a_{ij}, 定数 c_i を用いて、数値計算で近似解 x_j $(j=1,2,\cdots,n)$ を求め、その近似解 x_j $(j=1,2,\cdots,n)$ を n 元連立1次方程式(2)に代入して得られる残差が、(4)を満足する適当な誤差 ε_{ij}, ε_i を用いて

$$\sum_{j=1}^{n} a_{ij}\cdot x_j - c_i = -\sum_{j=1}^{n}\varepsilon_{ij}\cdot x_j + \varepsilon_i \qquad (7)$$

$$i=1,2,\cdots,n$$

とできるならば、近似解 x_j $(j=1,2,\cdots,n)$ は n 元連立1次方程式(2)を満足する近似解である。

解 x_j $(j=1,2,\cdots,n)$ が n 元連立1次方程式(2)を満足しているとの判定をする収束判定に、(4)を用いて、(7)の代りに不等式

$$\left|\sum_{j=1}^{n} a_{ij}\cdot x_j - c_i\right| \leqq \sum_{j=1}^{n}\Delta a_{ij}\cdot|x_j| + \Delta c_i \qquad (8)$$

$$i=1,2,\cdots,n$$

を用いるが、実際には、必ず(7)を満足する簡便な

$$\left|\sum_{j=1}^{n} a_{ij}\cdot x_j - C_i\right| \leqq \max_{j=1,2,\cdots,n}\left(\Delta a_{ij}\cdot|x_j|,\ \Delta C_i\right) \quad (9)$$
$$i=1,2,\cdots,n$$

を用いる。

Ⅲ　係数行列の数値的特異性

n元連立1次方程式(2)の係数間に、(4)を満足する適当なε_{ij}を用いると、

$$\sum_{j=1}^{n} a_{ij}\cdot\alpha_j = -\sum_{j=1}^{n}\varepsilon_{ij}\cdot\alpha_j \qquad i=1,2,\cdots,n \qquad (10)$$

の関係がなりたつ。少なくとも1つは0でない比例定数α_j($j=1,2,\cdots,n$)が存在すれば、n元連立1次方程式(2)は数値的に従属な式を含んでいる。

また(10)がなりたつと、

$$\sum_{j=1}^{n}(a_{ij}+\varepsilon_{ij})\cdot\alpha_j = 0 \qquad i=1,2,\cdots,n \qquad (10')$$

であり、$(a_{ij}+\varepsilon_{ij})$による係数行列は正則でないから

$$\sum_{i=1}^{n}(a_{ij}+\varepsilon_{ij})\cdot\beta_i = 0 \qquad j=1,2,\cdots,n \qquad (11')$$

$$\sum_{i=1}^{n} a_{ij}\cdot\beta_i = -\sum_{i=1}^{n}\varepsilon_{ij}\cdot\beta_i \qquad j=1,2,\cdots,n \qquad (11)$$

がなりたつ。したがって、n元連立1次方程式(2)が数値的に従属な式を含んでいるとの判定をするため、(4)を用いて、(10)，(11)の代りに

$$\left|\sum_{j=1}^{n} a_{ij}\cdot\alpha_j\right| \leqq \sum_{j=1}^{n}\Delta a_{ij}\cdot|\alpha_j| \qquad i=1,2,\cdots,n \qquad (12)$$

$$\left|\sum_{i=1}^{n} a_{ij}\cdot\beta_i\right| \leqq \sum_{i=1}^{n}\Delta a_{ij}\cdot|\beta_i| \qquad j=1,2,\cdots,n \qquad (13)$$

を用いる。

Ⅳ　ピボットと係数行列の特異性

ガウスの消去法を用いてn元連立1次方程式(2)の係数行列$[a_{ij}]$を左下3角行列$[l_{ij}]$$(i \geq j)$，右上3角行列$[m_{ij}]$$(i \leq j)$に分解すると，

$$l_{i1} = a_{i1} \qquad i = 1, 2, \cdots, n$$

$$m_{1j} = a_{1j} / l_{11} \qquad j = 1, 2, \cdots, n$$

$$l_{ij} = a_{ij} - \sum_{k=1}^{\min(i,j)-1} l_{ik} \cdot m_{kj} \qquad (i \geq j) \qquad \begin{cases} i = 2, 3, \cdots, n \\ j = 2, 3, \cdots, i \end{cases}$$

$$m_{ij} = \left(a_{ij} - \sum_{k=1}^{\min(i,j)-1} l_{ik} \cdot m_{kj} \right) / l_{ii} \qquad (i \leq j) \qquad \begin{cases} i = 2, 3, \cdots, n \\ j = i, i+1, \cdots, n \end{cases}$$

$$(\ m_{ii} = 1 \qquad i = 1, 2, \cdots, n\ ) \tag{14}$$

であるから，n元連立1次方程式(2)の係数a_{ij}は

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik} \cdot m_{kj} \tag{15}$$

である。

(12)，(13)を満足する列に関する比例定数$\alpha_j$$(j = 1, 2, \cdots, n)$，行に関する比例定数$\beta_i$$(i = 1, 2, \cdots, n)$が存在するとして，n元連立1次方程式(2)の係数行列をスケーリング[*1]し，次いで，完全枢軸選択のガウスの消去法で$(n-1)$回消去すると，それぞれ1行，1列を用いずに$\alpha_j$$(j = 1, 2, \cdots, n)$，$\beta_i$$(i = 1, 2, \cdots, n)$を求めることになる。（以下では，用いなかった1行，1列をそれぞれ第n行，第n列とする。）

$$\sum_{j=1}^{n} a_{ij} \cdot \alpha_j = 0 \qquad i = 1, 2, \cdots, n \qquad (16)$$

$$\sum_{i=1}^{n} a_{ij} \cdot \beta_i = 0 \qquad j = 1, 2, \cdots, n \qquad (16')$$

$(n-1)$回の枢軸選択によるガウスの前段消去を行なえば

$$\sum_{j=k}^{n} m_{kj} \cdot \alpha_j = 0 \qquad k = 1, 2, \cdots, n-1 \qquad (17)$$

$$\sum_{i=k}^{n} l_{ik} \cdot \beta_i = 0 \qquad k = 1, 2, \cdots, n-1 \qquad (17')$$

が得られるので、α_n, β_nをそれぞれ定数（例えば 1.0 ）とすれば、(17)，(17')からα_j（$j = 1, 2, \cdots, n-1$），β_i（$i = 1, 2, \cdots, n-1$）が求められる。したがって、(16)，(16')は(17)，(17')から

$$\sum_{j=1}^{n} a_{ij} \cdot \alpha_j = \sum_{k=1}^{i} l_{ik} \cdot \left(\sum_{j=k}^{n} m_{kj} \cdot \alpha_j \right) = 0 \qquad (18)$$

$$i = 1, 2, \cdots, n-1$$

$$\sum_{i=1}^{n} a_{ij} \cdot \beta_i = \sum_{k=1}^{j} \left(\sum_{i=k}^{n} l_{ik} \cdot \beta_i \right) \cdot m_{kj} = 0 \qquad (18')$$

$$j = 1, 2, \cdots, n-1$$

$$\sum_{j=1}^{n} a_{nj} \cdot \alpha_j = \sum_{k=1}^{n} l_{nk} \cdot \left(\sum_{j=k}^{n} m_{kj} \cdot \alpha_j \right) = l_{nn} \cdot m_{nn} \cdot \alpha_n \qquad (19)$$

$$\sum_{i=1}^{n} a_{in} \cdot \beta_i = \sum_{k=1}^{n} \left(\sum_{i=k}^{n} l_{ik} \cdot \beta_i \right) \cdot m_{kn} = l_{nn} \cdot m_{nn} \cdot \beta_n \qquad (19')$$

となる。すなわち、n元連立1次方程式(2)の係数行列は、第n式を除いて、(12)の條件をすべて満足しているので、

$$|l_{nn} \cdot \alpha_n| \leqq \sum_{j=1}^{n} \Delta a_{nj} \cdot |\alpha_j| \qquad m_{nn} = 1.0 \qquad (20)$$

を満足していれば、第n式（$i = n$）も(12)の條件を満足するので、n元連立1次方程式(2)は数値的に従属な式を含んでいることになる。

ガウスの消去法で用いるピボットの数値の積の絶対値 $|l_{11}\cdot l_{22}\cdot\cdots\cdot l_{nn}|$ は、ピボットの数値の採用の仕方に関係なく、係数行列の行列式の値の絶対値に等しく、一方、完全枢軸選択のガウスの消去法では、各消去段階でピボット l_{ii} ($i=1,2,\cdots,n$) として採用してよい係数の中で、絶対値最大の数値をピボットとして採用しているので、最後（n 回目）のピボット l_{nn} の絶対値は他のピボット $l_{11}, l_{22}, \cdots, l_{n-1\,n-1}$ の絶対値に比較して、一般に小さい。特に係数行列が悪条件の時にそれが著しい。

また n 元連立1次方程式(2)の係数行列をスケーリング*1し、次いで、完全枢軸選択のガウスの消去法で消去すると、$\alpha_n=1.0$, $\beta_n=1.0$ ならば

$$|\alpha_j|\leqq 1.0 \qquad |\beta_i|\leqq 1.0 \qquad i,j=1,2,\cdots,n-1$$

がほぼなりたつ。

*1 n 元連立1次方程式(2)の係数の絶対値が極端に異なる時は、次のようなスケーリングを行なう。

$$A_i=\max_{j=1,2,\cdots,n}(|a_{ij}|) \qquad i=1,2,\cdots,n$$

$$\bar{A}_j=\max_{i=1,2,\cdots,n}(|a_{ij}/A_i|) \qquad j=1,2,\cdots,n$$

$$a_{Nij}=a_{ij}/(A_i\cdot\bar{A}_j) \qquad i,j=1,2,\cdots,n$$

新しい係数行列 a_{Nij} の列に関する比例定数 α_{Nj} ($j=1,2,\cdots,n$)，行に関する比例定数 β_{Ni} ($i=1,2,\cdots,n$) から

$$\alpha_j = \alpha_{Nj} / \bar{A}_j \qquad j = 1, 2, \cdots, n$$

$$\beta_i = \beta_{Ni} / A_i \qquad i = 1, 2, \cdots, n$$

によって n 元連立1次方程式(2)の係数行列の列に関する比例定数 α_j ($j = 1, 2, \cdots, n$)，行に関する比例定数 β_i ($i = 1, 2, \cdots, n$) を求める．

條件(4)を満足する，すべての ε_{ij} ($i, j = 1, 2, \cdots, n$) を考慮しているわけではないので，條件(20)を満足しなくても，條件(12)を満足する場合がある．

そこで，あまり実用にはならないが，(12)を満足しない場合を求めてみる．まず $\alpha_n = 1.0$ と(17)から α_j ($j = 1, 2, \cdots, n$) を求め，次のようなスケーリングをすると，n 元連立1次方程式(2)の係数は

$$\sum_{j=1}^{n} \bar{a}_{ij} \cdot \bar{\alpha}_j = 0 \qquad i = 1, 2, \cdots, n \qquad (21)$$

$$\bar{a}_{ij} = (a_{ij} \cdot \alpha_j) / \left(\sum_{k=1}^{n} \Delta a_{ik} \cdot |\alpha_k|\right)$$

$$\Delta \bar{a}_{ij} = (\Delta a_{ij} \cdot \alpha_j) / \left(\sum_{k=1}^{n} \Delta a_{ik} \cdot |\alpha_k|\right)$$

となり，n 元連立1次方程式(2)が数値的に従属な式を含んでいるとの判定をするため，(12)の代りに

$$\left|\sum_{j=1}^{n} \bar{a}_{ij} \cdot \bar{\alpha}_j\right| \leqq \sum_{j=1}^{n} \Delta \bar{a}_{ij} \cdot |\bar{\alpha}_j| \qquad (22)$$

$$i = 1, 2, \cdots, n$$

を用いる．$\bar{\alpha}_n = 1.0$ とすると，

$$\bar{\alpha}_j \fallingdotseq 1.0 \qquad j = 1, 2, \cdots, n-1$$

であり、ほぼ次式がなりたつ。

$$\left| \sum_{j=1}^{n} \bar{a}_{ij} \cdot \bar{\alpha}_j \right| \leqq 1.0 \qquad i = 1, 2, \cdots, n \qquad (24)$$

次いで、完全枢軸選択のガウスの消去法で係数行列の $\bar{a}_{ij}$ を $\bar{l}_{ij}$ $(i \geqq j)$, $\bar{m}_{ij}$ $(i \leqq j)$ に分解し、(17), (18), (19) と同様に

$$\sum_{j=k}^{n} \bar{m}_{kj} \cdot \bar{\alpha}_j = 0 \qquad k = 1, 2, \cdots, n-1 \qquad (25)$$

$$\sum_{j=1}^{n} \bar{a}_{ij} \cdot \bar{\alpha}_j = \sum_{k=1}^{i} \bar{l}_{ik} \cdot \left(\sum_{j=k}^{n} \bar{m}_{kj} \cdot \bar{\alpha}_j \right) = 0 \qquad (26)$$

$$i = 1, 2, \cdots, n-1$$

$$\sum_{j=1}^{n} \bar{a}_{nj} \cdot \bar{\alpha}_j = \sum_{k=1}^{n} \bar{l}_{nk} \cdot \left(\sum_{j=k}^{n} \bar{m}_{kj} \cdot \bar{\alpha}_j \right) = \bar{l}_{nn} \cdot \bar{m}_{nn} \cdot \bar{\alpha}_n \qquad (27)$$

$$(\bar{m}_{nn} = 1.0)$$

となる。n 元連立1次方程式(21)の第 n 式を除いて、(22)の條件をすべて満足しているので

$$\left| \bar{l}_{nn} \cdot \bar{\alpha}_n \right| \leqq \sum_{j=1}^{n} \Delta \bar{a}_{nj} \cdot \left| \bar{\alpha}_j \right| \qquad (28)$$

を満足していれば、n 元連立1次方程式(2)の第 n 式も(20)の條件を満足するので、n 元連立1次方程式(2)は数値的に従属な式を含んでいることになる。

n 元連立1次方程式(21)の第 n 式を除いて、(26)では、残差を0としているが、第 i 式 $(i \neq n)$ の残差を(22)の右辺

$$\sum_{j=1}^{n} \Delta \bar{a}_{ij} \cdot \left| \bar{\alpha}_j \right| \qquad (\alpha_n = 1.0) \qquad (29)$$

$$i = 1, 2, \cdots, n-1$$

とすると、比例定数 $\bar{\alpha}_j$ $(j = 1, 2, \cdots, n-1)$ の変化分は

$$b_{ji}\cdot\left(\sum_{k=1}^{n}\Delta\bar{a}_{ik}\cdot|\bar{\alpha}_k|\right) \qquad (30)$$

$$i,j=1,2,\cdots\cdots,n-1$$

（ただし、係数行列 $[b_{ij}]$ は係数行列 $[a_{ij}]$（$i,j=1,2,\cdots\cdots,n-1$）の逆行列）

となる。したがって、(22)の第1式から第（$n-1$）式までの右辺を、それぞれ残差とすると、比例定数 $\bar{\alpha}_j$（$j=1,2,\cdots\cdots,n-1$）の変化分 $\Delta\bar{\alpha}_j$（$j=1,2,\cdots\cdots,n-1$）は

$$\Delta\bar{\alpha}_j=\sum_{i=1}^{n-1}b_{ji}\cdot\left(\sum_{k=1}^{n}\Delta\bar{a}_{ik}\cdot|\bar{\alpha}_k|\right) \qquad (31)$$

$$j=1,2,\cdots\cdots,n-1$$

$$\Delta\bar{\alpha}_n=0\quad,\qquad\bar{\alpha}_n=1.0$$

となる。ここで $\bar{\alpha}_j+\Delta\bar{\alpha}_j$（$j=1,2,\cdots\cdots,n$）を n 元連立1次方程式(21)の第 n 式に代入する。

$$\begin{aligned}&\sum_{j=1}^{n}\bar{a}_{nj}\cdot(\bar{\alpha}_j+\Delta\bar{\alpha}_j)\\&=\bar{l}_{nn}\cdot\bar{m}_{nn}\cdot\bar{\alpha}_n+\sum_{j=1}^{n-1}\bar{a}_{nj}\cdot\left\{\sum_{i=1}^{n-1}b_{ji}\cdot\left(\sum_{k=1}^{n}\Delta\bar{a}_{ik}\cdot|\bar{\alpha}_k|\right)\right\}\\&=\bar{l}_{nn}+\sum_{i=1}^{n-1}\left\{\left(\sum_{j=1}^{n-1}\bar{a}_{nj}\cdot b_{ji}\right)\cdot\left(\sum_{k=1}^{n}\Delta\bar{a}_{ik}\cdot|\bar{\alpha}_k|\right)\right\}\end{aligned} \qquad (32)$$

$$\bar{m}_{nn}=1.0\quad,\qquad\bar{\alpha}_n=1.0$$

(32)より、次式を満足すれば、n 元連立1次方程式(2)は数値的に従属な式を含まないとみなしている。

$$\begin{aligned}|\bar{l}_{nn}\cdot\bar{m}_{nn}\cdot\bar{\alpha}_n|>&\sum_{i=1}^{n-1}\left\{\left(\left|\sum_{j=1}^{n-1}\bar{a}_{nj}\cdot b_{ji}\right|\right)\cdot\left(\sum_{k=1}^{n}\Delta\bar{a}_{ik}\cdot|\bar{\alpha}_k|\right)\right\}\\&+\sum_{j=1}^{n}\Delta\bar{a}_{nj}\cdot|\bar{\alpha}_j+\Delta\bar{\alpha}_j|\end{aligned} \qquad (33)$$

ただし、(33)の中に含まれる $\Delta\alpha_j$ は次式で示される。

$$\Delta\bar{\alpha}_j = \sum_{i=1}^{n-1} \bar{b}_{ji} \cdot \left(\sum_{k=1}^{n} \Delta\bar{a}_{ik} \cdot |\bar{\alpha}_k| \right) \qquad j = 1, 2, \cdots, n-1 \qquad (34)$$

$$\bar{b}_{ji} = b_{ji} \quad , \quad \left(\sum_{k=1}^{n-1} \bar{a}_{nk} \cdot b_{ki} \right) \geq 0$$

$$\bar{b}_{ji} = -b_{ji} \quad , \quad \left(\sum_{k=1}^{n-1} \bar{a}_{nk} \cdot b_{ki} \right) < 0$$

しかし、比例定数$\bar{\alpha}_j$が$(\bar{\alpha}_j + \Delta\bar{\alpha}_j)$に変化したので、條件(22)は

$$\left| \sum_{j=1}^{n} \bar{a}_{ij} \cdot (\bar{\alpha}_j + \Delta\bar{\alpha}_j) \right| \leqq \sum_{j=1}^{n} \Delta\bar{a}_{ij} \cdot |\bar{\alpha}_j + \Delta\bar{\alpha}_j| \qquad (35)$$

$$i = 1, 2, \cdots, n$$

となる。したがって、(33)の條件でn元連立1次方程式(2)が数値的に従属な式を含まないとは、厳密にはいえないが、本来の誤差の性質から考えて、実用的には充分である。

ここでn元連立1次方程式(2)の誤差を含む係数行列の條件数として

$$\rho = \min \frac{\left| \sum_{i=1}^{n} \beta_i \cdot \left(\sum_{j=1}^{n} a_{ij} \cdot \alpha_j \right) \right|}{\sum_{i=1}^{n} |\beta_i| \cdot \left(\sum_{j=1}^{n} \Delta a_{ij} \cdot |\alpha_j| \right)} \qquad (36)$$

（ただし、係数行列$[a_{ij}]$は正則であり、且つ比例定数α_j, β_i $(i, j = 1, 2, \cdots, n)$はそれぞれ少なくとも1つは0でない。）

を定義する。この條件数が

$$\rho \gg 1.0$$

であれば、誤差を含む係数行列の性質が良く、ρが1に近ければ近い程、この係数行列を持つn元連立1次方程式(2)は

式の間の独立性を数値的にだんだん失ってゆく。

Ⅴ　枢軸（ピボット）と固有値

n元連立1次方程式(2)の係数をスケーリング*1し、次いで、完全枢軸選択によるガウスの消去法で計算したとき、計算途中において枢軸（ピボット）の数値に桁落ち誤差が入る場合、n元連立1次方程式(2)の近似解を求めるとともに、n元連立1次方程式(2)の係数行列$[a_{ij}]$をスケーリング*1して得られる係数行列の列および行の比例定数$\alpha_j^{(0)}$, $\beta_i^{(0)}$ ($i,j=1,2,\cdots\cdots,n$)を(16), (16')を用いて求める。

この比例定数$\alpha_j^{(0)}$, $\beta_i^{(0)}$ ($i,j=1,2,\cdots\cdots,n$)とスケーリング*1して得られる係数行列とで

$$\left|\sum_{j=1}^{n} a_{ij}^{(0)}\cdot\alpha_j^{(0)}\right| \qquad i=1,2,\cdots\cdots,n \qquad (37)$$

$$\left|\sum_{i=1}^{n} a_{ij}^{(0)}\cdot\beta_i^{(0)}\right| \qquad j=1,2,\cdots\cdots,n \qquad (37')$$

（ただし、$a_{ij}^{(0)}$はn元連立1次方程式(2)の係数行列をスケーリング*1して得られる係数行列）

を計算すれば、(18), (18'), (19), (19')と枢軸（ピボット）が桁落ちしていることから

$$\left|\sum_{j=1}^{n} a_{ij}^{(0)}\cdot\alpha_j^{(0)}\right| \fallingdotseq 0 \qquad i=1,2,\cdots\cdots,n-1 \qquad (38)$$

$$\left|\sum_{i=1}^{n} a_{ij}^{(0)}\cdot\beta_i^{(0)}\right| \fallingdotseq 0 \qquad j=1,2,\cdots\cdots,n-1 \qquad (38')$$

$$\left|\sum_{j=1}^{n} a_{nj}^{(0)}\cdot\alpha_j^{(0)}\right| = \left|l_{nn}^{(0)}\cdot m_{nn}^{(0)}\cdot\alpha_n^{(0)}\right| \ll \max_{j=1,2,\cdots,n}\left(\left|a_{nj}^{(0)}\cdot\alpha_j^{(0)}\right|\right) \qquad (39)$$

$$\left|\sum_{i=1}^{n} a_{in}^{(0)}\cdot\beta_i^{(0)}\right| = \left|l_{nn}^{(0)}\cdot m_{nn}^{(0)}\cdot\beta_n^{(0)}\right| \ll \max_{i=1,2,\cdots,n}\left(\left|a_{in}^{(0)}\cdot\beta_i^{(0)}\right|\right) \qquad (39')$$

となる。すなわち、この比例定数 $\alpha_j^{(0)}$, $\beta_i^{(0)}$ $(i,j=1,2,\cdots\cdots,n)$ は、一般に、係数行列 $[a_{ij}^{(0)}]$ の含む固有値の中で絶対値最小の固有値の固有ベクトルの近似ベクトルである。したがってこの場合、n 元連立1次方程式(2)の係数行列 $[a_{ij}]$ をスケーリング*1して得られる係数行列 $[a_{ij}^{(0)}]$ の絶対値最大の固有値 $\lambda_{max}^{(0)}$ と絶対値最小の固有値 $\lambda_{min}^{(0)}$ との関係が

$$\left|\lambda_{max}^{(0)} / \lambda_{min}^{(0)}\right| \gg 1.0 \qquad (40)$$

である。しかし、スケーリング*1されていない一般の係数行列の場合は、(40)を満足しているからといって、(38), (38′), (39), (39′)を満足するとは限らない。

比例定数 $\alpha_j^{(0)}$, $\beta_i^{(0)}$ $(i,j=1,2,\cdots\cdots,n)$ を対応する固有ベクトルの近似ベクトルとした場合の精度の桁数は、(38), (38′), (39), (39′)から、それぞれ

$$log_t\left\{\max_{j=1,2,\cdots,n}\left(\left|a_{nj}^{(0)}\cdot\alpha_j^{(0)}\right|\right) / \left|l_{nn}^{(0)}\cdot m_{nn}^{(0)}\cdot\alpha_n^{(0)}\right|\right\} \qquad (41)$$

$$log_t\left\{\max_{i=1,2,\cdots,n}\left(\left|a_{in}^{(0)}\cdot\beta_i^{(0)}\right|\right) / \left|l_{nn}^{(0)}\cdot m_{nn}^{(0)}\cdot\beta_n^{(0)}\right|\right\} \qquad (41')$$

ただし、t 進法

の程度である。

次に、n 元連立1次方程式(2)の係数行列 $[a_{ij}]$ をスケーリング*1して得られる係数行列 $[a_{ij}^{(0)}]$ の比例定数 $\alpha_i^{(0)}$, $\beta_i^{(0)}$ $(i,j=1,2,\cdots\cdots,n)$ に対応する固有値、固有ベクトルをそれぞれ $\lambda_T^{(0)}$, $\alpha_{Tj}^{(0)}$, $\beta_{Ti}^{(0)}$ $(i,j=1,2,\cdots\cdots,n)$ とすると

$$\sum_{i=1}^{n}\beta_{Ti}^{(0)}\cdot\left(\sum_{j=1}^{n}a_{ij}^{(0)}\cdot\alpha_{Tj}^{(0)}\right)=\lambda_T^{(0)}\cdot\sum_{i=1}^{n}\beta_{Ti}^{(0)}\cdot\alpha_{Ti}^{(0)} \qquad (42)$$

であり、この計算での桁落ち誤差の入る桁数はほぼ

$$P_T = \log_t \left\{ \frac{\sum_{i=1}^{n} \left| \beta_{Ti}^{(0)} \right| \cdot \left(\sum_{j=1}^{n} \left| a_{ij}^{(0)} \cdot \alpha_{Tj}^{(0)} \right| \right)}{\left| \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \right|} \right\} \tag{43}$$

$$P_T > 1 \qquad t : t\text{進数}$$

である。また、$\alpha_j^{(0)}, \beta_i^{(0)}$ $(i, j = 1, 2, \cdots\cdots, n)$ の誤差の入っていない桁数をそれぞれ

$$P_\alpha = \min_{j=1,2,\cdots,n} \log_t \left\{ \left| \alpha_{Tj}^{(0)} / \Delta\alpha_{Tj}^{(0)} \right| \right\}, \qquad \alpha_{Tj}^{(0)} = \alpha_j^{(0)} + \Delta\alpha_{Tj}^{(0)}$$

$$P_\alpha > 1 \tag{44}$$

$$P_\beta = \min_{i=1,2,\cdots,n} \log_t \left\{ \left| \beta_{Ti}^{(0)} / \Delta\beta_{Ti}^{(0)} \right| \right\}, \qquad \beta_{Ti}^{(0)} = \beta_i^{(0)} + \Delta\beta_{Ti}^{(0)}$$

$$P_\beta > 1 \tag{45}$$

とすると

$$\sum_{i=1}^{n} \beta_i^{(0)} \cdot \left(\sum_{j=1}^{n} a_{ij}^{(0)} \cdot \alpha_j^{(0)} \right) = \sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \left(\sum_{j=1}^{n} a_{ij}^{(0)} \cdot \alpha_{Tj}^{(0)} \right) - \lambda_T^{(0)} \cdot \sum_{j=1}^{n} \beta_{Tj}^{(0)} \cdot \Delta\alpha_{Tj}^{(0)}$$
$$- \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \Delta\beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} + \sum_{i=1}^{n} \Delta\beta_{Ti}^{(0)} \cdot \left(\sum_{j=1}^{n} a_{ij}^{(0)} \cdot \Delta\alpha_{Tj}^{(0)} \right)$$

$$\sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \left(\sum_{j=1}^{n} a_{ij}^{(0)} \cdot \alpha_{Tj}^{(0)} \right) = \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \tag{46}$$

の式の中では、(44), (45) から

$$\left| \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \right| \gg \left| \lambda_T^{(0)} \cdot \sum_{j=1}^{n} \beta_{Tj}^{(0)} \cdot \Delta\alpha_{Tj}^{(0)} \right|$$
$$\left| \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \right| \gg \left| \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \Delta\beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \right| \tag{47}$$

なる関係があり

$$\left| \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \right| \gg \left| \sum_{i=1}^{n} \Delta\beta_{Ti}^{(0)} \cdot \left(\sum_{j=1}^{n} a_{ij}^{(0)} \cdot \Delta\alpha_{Tj}^{(0)} \right) \right| \tag{48}$$

を満足するならば、すなわち、(43), (44), (45) から

$$P_T < P_\alpha + P_\beta \tag{49}$$

であるならば、

$$\sum_{i=1}^{n} \beta_i^{(0)} \cdot \left(\sum_{j=1}^{n} a_{ij}^{(0)} \cdot \alpha_j^{(0)} \right) \fallingdotseq \lambda_T^{(0)} \cdot \sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \qquad (50)$$

である。したがって、固有値は

$$\lambda_T^{(0)} \fallingdotseq \sum_{i=1}^{n} \beta_i^{(0)} \cdot \left(\sum_{j=1}^{n} a_{ij}^{(0)} \cdot \alpha_j^{(0)} \right) \Big/ \left(\sum_{i=1}^{n} \beta_{Ti}^{(0)} \cdot \alpha_{Ti}^{(0)} \right)$$

$$\fallingdotseq l_{nn}^{(0)} \cdot m_{nn}^{(0)} \cdot \beta_n^{(0)} \cdot \alpha_n^{(0)} \Big/ \left(\sum_{i=1}^{n} \beta_i^{(0)} \cdot \alpha_i^{(0)} \right) \qquad (51)$$

で求められる。

ここで、係数行列 $[a_{ij}^{(0)}]$ の最後の列を

$$a_{in}^{(0)} \longrightarrow 10^{-m} \cdot a_{in}^{(0)} \qquad m \gg 1 \qquad (52)$$

$$i = 1, 2, \cdots, n$$

とすると、固有値は

$$\lambda_T^{(0)} \fallingdotseq \frac{\sum_{i=1}^{n} \beta_i^{(0)} \cdot \left\{ \sum_{j=1}^{n-1} a_{ij}^{(0)} \cdot \alpha_j^{(0)} + \left(10^{-m} \cdot a_{in}^{(0)} \right) \cdot \left(10^{m} \cdot \alpha_n^{(0)} \right) \right\}}{\sum_{i=1}^{n-1} \beta_i^{(0)} \cdot \alpha_i^{(0)} + \beta_n^{(0)} \cdot \left(10^{m} \cdot \alpha_n^{(0)} \right)} \qquad (53)$$

となり、λ_T の絶対値は小さくなる。しかし、係数行列の條件数(36)は変化したわけではない。すなわち、(40)の條件が満足されても、係数行列の性質が必ずしも悪いわけではなく、(40)の條件を満足している場合、係数行列の性質が悪いとはいえるのは、スケーリング*1された係数行列の場合である

また、次式を満足する固有値 λ_T は誤差のみである。

$$\left| \lambda_T \cdot \alpha_{Ti} \right| \leqq \left| \sum_{j=1}^{n} \Delta a_{ij} \cdot \alpha_{Tj} \right| , \quad i = 1, 2, \cdots, n \qquad (54)$$

ただし、λ_T は n 元連立1次方程式(2)の係数行列の固有値であり、α_{Ti} $(i = 1, 2, \cdots, n)$ は固有値 λ_T に対応

する固有ベクトルである.

Ⅵ 係数および定数の誤差と近似解の誤差

n 元連立1次方程式 (2) の係数行列 $[a_{ij}]$ の逆行列を

$$A^{-1} = \bar{\bar{B}} = \left[\quad \bar{\bar{b}}_{ij} \quad \right] \tag{55}$$

とすると、n 元連立1次方程式 (2) の解 x_j ($j=1,2,\cdots,n$) の誤差は、ほぼ

$$\Delta x_{Tj} \fallingdotseq \sum_{i=1}^{n} \bar{\bar{b}}_{ji} \cdot \delta_{Ti} \qquad j=1,2,\cdots,n \tag{56}$$

$$\delta_{Ti} = -\sum_{j=1}^{n} \varepsilon_{Tij} \cdot x_j + \varepsilon_{Ti} \qquad i=1,2,\cdots,n$$

である。したがって

$$-\sum_{i=1}^{n} |\bar{\bar{b}}_{ji}| \cdot \Delta d_{Ti} \leqq \Delta x_{Tj} \leqq \sum_{i=1}^{n} |\bar{\bar{b}}_{ji}| \cdot \Delta d_{Ti} \tag{57}$$

$$\Delta d_{Ti} = \sum_{j=1}^{n} \Delta a_{ij} \cdot |x_j| + \Delta C_i \qquad j=1,2,\cdots,n$$

であるが、実用としては

$$-\max_{i=1,2,\cdots,n} \left(|\bar{\bar{b}}_{ji}| \cdot \Delta d_i \right) \leqq \Delta x_j \leqq \max_{i=1,2,\cdots,n} \left(|\bar{\bar{b}}_{ji}| \cdot \Delta d_i \right)$$

$$\Delta d_i = \max_{j=1,2,\cdots,n} \left(\Delta a_{ij} \cdot |x_j| , \Delta C_i \right) \tag{58}$$

$$j=1,2,\cdots,n$$

を用いている。

また、(56) からわかるように、解 x_j ($j=1,2,\cdots,n$) の誤差 Δx_{Tj} の範囲を (57), (58) で示したが、すべての解が同時に最大の誤差を含むわけではない。

Ⅶ 補足

今回は、n元連立1次方程式(2)の係数a_{ij}, 定数C_iが含む誤差ε_{ij}, ε_iはそれぞれ独立であるとして考えてきたが、一般に、独立でないときには、今回の解の誤差より解の誤差が小さくなる。また、誤差ε_{ij}, ε_iが独立のときの例は参照文献①, ～, ④にある。

n元連立1次方程式(2)を完全枢軸選択によるガウスの消去法で計算する場合、必要以上の計算桁数をとらないで、

$$\log_t \left\{ \max_{i,j=1,2,\cdots,n} \left(\left| a_{ij} / \Delta a_{ij} \right| \right) \right\} \qquad (59)$$

t : t進数

に、丸めの誤差を防ぐために必要な桁を数桁加えた桁数で計算するためには、スケーリング[*1]の方法として

$$a_{ij}^{(0)} = \left(a_{ij} \cdot x_j \right) / \max_{j=1,2,\cdots,n} \left(\left| \Delta a_{ij} \cdot x_j \right|, \Delta C_i \right) \qquad (60)$$

$$i, j = 1, 2, \cdots, n$$

（ただし、x_j ($j=1,2,\cdots,n$)はn元連立1次方程式(2)の解

が好ましい。しかし、x_j ($j=1,2\cdots,n$)の値は未知であるので、近似解が既知の場合には、その近似解[*2]を用いる。

*2 (60)の分母で選ばれる項

$$\Delta a_{iM} \cdot x_{iM} = \max_{j=1,2,\cdots,n} \left(\left| \Delta a_{ij} \cdot x_j \right|, \Delta C_i \right) \qquad (61)$$

$$i = 1, 2, \cdots, n$$

に含まれる解x_{iM} ($i=1,2,\cdots,n$)およびピボットと

して採用される係数を持つ解は、少なくとも1～2桁の精度を必要とする。

最後に、ピボットに関することであるが、スケーリング*1をした後、完全枢軸選択によるガウスの消去法を用いる場合、ピボットの積 $|l_{11}^{(0)} \cdot l_{22}^{(0)} \cdots\cdots \cdot l_{nn}^{(0)}|$ が係数行列の行列式の値の絶対値に等しく、ピボットの選び方とは関係なく一定であるから、ピボットとして用いてよい係数の中で絶対値最大の数値を用いてゆくと、前段消去の最後のほうで採用するピボットの数値の絶対値が小さくなり、誤差が入る。したがって、「ピボットとして用いてよい係数の中で絶対値最大の数値をピボットに用いず、絶対値の大きい順に数えて数番目の数値をピボットとして採用することが好ましい。」との考えもあるが、この考えは、n元連立1次方程式(2)の係数 a_{ij} に誤差を考え、数値的に従属な式を含むか否かを考えれば、今回述べたことから、意味がないことがわかろう。

参照文献

①「数値の誤差と数式の誤差」　平野菅保
日本物理学会　応用数学力学講演会（1965年）
②「代数方程式の解法及び誤差」　平野菅保
情報処理学会
第8回プログラミング・シンポジューム（1967年）

③「関数近似と誤差の問題」　平野菅保

電子計算機のための数値計算法Ⅱ　　培風館

山内二郎、森口繁一、一松信　共編（1967年）

④「INPUT の精度（けた数）で行う計算」　平野菅保

技術者のための電子演算手法　　コロナ社

木村久男編　　（1970年）

⑤「多元連立1次方程式の精密な解き方」　平野菅保

コンピューターによる構造工学講座 Ⅱ－1－A

戸川隼人，藤井宏，三好哲彦，平野菅保　共著

培風館　（1971年）

⑥「連立方程式の数値解の一つの試み」　平野菅保

数理解析研究所講究録 107

京都大学数理解析研究所　（1970年）